เรื่องน่าอ่าน
ทศเทพ บุญสิลา

# ข้อมูลข่าวสารในระบบจัดเก็บและค้นหา

***ใช้ระบบจัดเก็บข้อมูลในรูปแฟ้มข้อมูลเพื่อความสะดวกและรวดเร็วในการค้นหา***

ปัจจุบันนี้ถือเป็นยุคแห่งข่าวสารข้อมูล (Information Age) ซึ่งเป็นผลพลอยได้มาจากการที่มีการพัฒนาวิทยาการ ไม่ว่าจะเป็นเศรษฐกิจ วิทยาศาสตร์ การเมือง การสื่อสารได้รับการพัฒนาให้ล้ำยุคในรูปแบบต่าง ๆ ข่าวสารข้อมูลก็เพิ่มมากขึ้นตามไปด้วย การที่จะนำข้อมูลข่าวสารจำนวนมากใช้ให้เป็นประโยชน์ และมีประสิทธิภาพจะต้องมีวิธีการที่ชาญฉลาด จึงจะได้ประโยชน์สูงสุด มิฉะนั้นก็จะเหมือนงมเข็มในมหาสมุทร

ระบบจัดเก็บและค้นหาข้อมูลข่าวสาร (Information Storage and Retrieval System) จึงถูกนำมาใช้ในการจัดการกับข้อมูลข่าวสารจำนวนมากนี้ เพื่อให้เป็นเครื่องกรองหรือเครื่องคัดเลือกข้อมูลข่าวสารแทนเรา และช่วยในการค้นหาเอกสารหรือหนังสือที่เกี่ยวข้องกับความต้องการแก่ผู้ใช้ คือระบบจัดการนี้จะดำเนินการจัดเก็บข้อมูลข่าวสารซึ่งอยู่ในรูปแฟ้มข้อมูล (Data-file) กลั่นกรองและทำการสร้างดัชนีเพื่อจัดโครงสร้างในการเรียกกลับมาใช้ได้รวดเร็วและตรงกับความต้องการของผู้ใช้มากที่สุด

ระบบจัดเก็บและค้นหาข้อมูลข่าวสารในปัจจุบันได้ถูกพัฒนาให้ใช้กับการเก็บเอกสารของบริษัทและองค์กรต่าง ๆ ในต่างประเทศเช่น The Wall Street Journal และ Library of congress ของสหรัฐอเมริกา ISRS จึงเป็นระบบจัดการข่าวสารที่น่าจับตามอง และน่าจะมีบทบาทสำคัญในการจัดการข้อมูลข่าวสาร (ซึ่งมีจำนวนมาก) ในอนาคตอันใกล้

## ระบบในปัจจุบัน

เนื่องจากมีการติดต่อสื่อสารกันระหว่างเครื่องคอมพิวเตอร์เพิ่มมากขึ้น และเครื่องคอมพิวเตอร์ก็มีประสิทธิภาพมากขึ้น จึงทำให้ระบบจัดการข่าวสารข้อมูล มีความสำคัญต่อการทำงานในยุคนี้ การเรียกใช้ข่าวสารอาจทำโดยออนไลน์หรือซีดีรอมก็ได้ เนื่องจากข่าวสารข้อมูลนั้นมีความสำคัญเพิ่มมากขึ้น การมีข่าวสารข้อมูลเก็บไว้จำนวนมาก ไม่ได้หมายความว่าจะได้รับประโยชน์จากข่าวสารข้อมูลนั้นมากขึ้น ถ้าหากว่าการค้นหาเอกสารที่ต้องการนั้นกระทำได้ยากและใช้เวลานาน

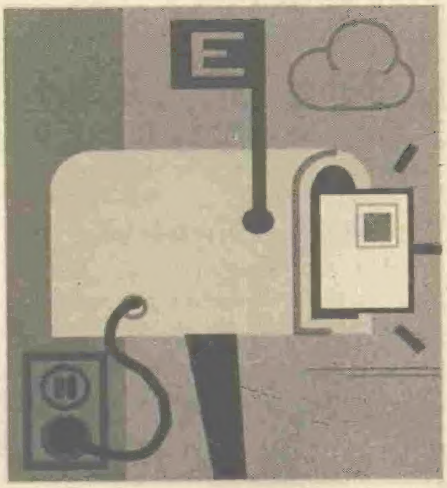
การจัดเก็บข้อมูลข่าวสารจำนวนมากจะต้องกระทำเป็นระบบ เพื่อให้ง่าย

ต่อการเรียกใช้หรือค้นหา การจัดเก็บข้อมูลข่าวสารที่มีรูปแบบต่าง ๆ กัน สไตล์การเขียนที่ต่าง ๆ กัน จากนักเขียนหลายคน ทำให้ลักษณะของข้อมูลข่าวสารไม่มีรูปแบบที่แน่ชัด (Free text information) ลักษณะของระบบข่าวสารข้อมูล (Information System) นี้เองที่แตกต่างจากระบบจัดการฐานข้อมูล (Database Management System) ซึ่งเป็นที่รู้จักแพร่หลายในปัจจุบัน การเก็บของข้อมูลของระบบฐานข้อมูลกระทำได้ง่าย เพราะโครงสร้างของข้อมูลมีรูปแบบและโครงสร้างที่กำหนดไว้อย่างชัดเจน เพราะเหตุนี้การค้นหาข้อมูลก็ทำได้ง่ายด้วย เทคโนโลยีซึ่งได้รับการพัฒนาขึ้นมาใช้จัดการกับระบบฐานข้อมูล จึงไม่สามารถนำมาใช้กับระบบข่าวสารข้อมูลได้

## ผู้ใช้

ผู้คนส่วนมากอาจเผชิญกับความต้องการข้อมูลข่าวสารในบางครั้ง เมื่อเกิดความจำเป็นที่จะต้องได้มาซึ่งข้อมูลข่าวสารนั้น ๆ บางคนก็อาจจะหันไปถามเพื่อน หรือไปค้นหาในห้องสมุด หาจากหนังสือ เอกสาร หรือสิ่งพิมพ์ต่าง ๆ หลายครั้งที่ค้นหาข้อมูลข่าวสารแล้วไม่พบ อาจจะได้คำตอบจากเพื่อน ๆ ว่า "ไม่รู้ซิ" หรือค้นหนังสือในห้องสมุดทั้งวันแล้วได้ข้อมูลข่าวสารไม่คุ้มกับเวลาที่เสียไป ระบบข้อมูลข่าวสารนั้นมีลักษณะคล้ายกับดัชนีในห้องสมุดนั้นเอง

โดยทั่วไปการค้นหาหนังสือหรือเอกสารที่เราต้องการ ก็จะไปค้นหาดัชนีเพื่อจะรู้ว่าเอกสารที่เราต้องการถูกเก็บไว้ที่ไหน ในการค้นหาอาจจะต้องเดินกลับไปกลับมาหลายครั้ง และต้องอ่านหนังสือหลายเล่มจึงจะได้ข้อมูลที่ต้องการ หากมีการนำระบบข้อมูลข่าวสารมาใช้ก็จะสามารถลดเวลาในการค้นหาเอกสารข้อมูลที่ต้องการ การใช้ระบบข้อมูลข่าวสารในการค้นหาข้อมูลนั้นไม่ได้จำกัดอยู่เพียงในห้องสมุด บริษัท ห้างร้านก็สามารถนำระบบนี้ไปใช้ได้เพื่อเพิ่มประสิทธิภาพในการทำงานให้แก่องค์กรนั้น ๆ สำนักงานทนายความอาจจะใช้ระบบข้อมูลข่าวสารในการค้นหากฎหมายมาตราต่าง ๆ หรือเอกสารตีความที่มีลักษณะคล้ายกับที่ทนายผู้นั้นกำลังว่าความเพื่อหาแนวทาง สำนักพิมพ์อาจมีความจำเป็นในการเก็บบทความที่พิมพ์ไปแล้วและสามารถค้นหาได้ง่าย ในขณะที่เราก้าวไปสู่ยุคแห่งข้อมูลข่าวสาร ระบบจัดการข้อมูลข่าวสารก็ยิ่งมีความสำคัญมากยิ่งขึ้นเป็นเงาตามตัว



## การจัดการ

ในโลกแห่งข้อมูลข่าวสารที่มีข้อมูลข่าวสารเพิ่มขึ้นอย่างมหาศาลทุก ๆ นาทีการจัดการกับข้อมูลข่าวสารต่าง ๆ จะต้องกระทำอย่างรวดเร็วมิฉะนั้นจะไม่ทันการณ์ ขั้นตอนในการจัดการกับข้อมูลนั้นแรกสุด คือ การถ่ายทอดข้อมูลข่าวสารจากเอกสารสิ่งพิมพ์ให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถเรียกใช้ได้ (แฟ้มข้อมูลต่าง ๆ) (Data Entry) ขั้นตอนนี้ไม่ยาก อาจทำได้โดยการพิมพ์ด้วยมือหรือใช้สแกนเนอร์ก็ได้

การทำงานของคอมพิวเตอร์นั้นแตกต่างจากมนุษย์เรามาก รูปแบบของไฟล์ข้อมูลข่าวสารจะต้องถูกจัดให้เป็นแบบมาตรฐาน (Normalizing) การจัดรูปแบบนี้ข้อมูลจะถูกนำมาแยกเป็นส่วน ๆ อาจจะเป็นชื่อผู้แต่ง วันที่พิมพ์ ชื่อเรื่อง เนื้อหา ฯลฯ เพื่อให้ง่ายต่อการจัดการเก็บและการทำโครงสร้างในขั้นตอนต่อไป ขั้นตอนต่อมาก็จะเป็นการทำดัชนี (Indexing) เพื่อใช้ในการอ้างอิง การทำดัชนีเรื่องในแบบดั้งเดิมนั้นจะต้องใช้คนเป็นผู้อ่านบทความหรือหนังสือนั้น ๆ แล้วสรุปเนื้อหาของบทความว่ามีความเกี่ยวข้องกับเรื่องใดบ้าง จึงเลือกคำดัชนีที่ใช้แทนความหมายของเนื้อหาของบทความนั้น ๆ ให้เหมาะสม ในการทำดัชนีโดยคนนั้นจะต้องใช้เวลาในการอ่านและเลือกคำดัชนี (Indexword) นาน แต่มีข้อดีที่ได้คำดัชนีที่กระชับตรงกับความหมายของเนื้อหา

เพื่อให้การทำดัชนีรวดเร็ว จึงใช้เครื่องคอมพิวเตอร์ทำแทน ในการทำดัชนีแบบอัตโนมัตินี้มีข้อสมมติฐานที่ว่าความหมายหรือหัวใจของเนื้อหานั้นมีความสัมพันธ์กับคำที่ปรากฏในบทความในระดับหนึ่งคือว่า คำที่ปรากฏในบทความบ่อยครั้งเกินไป อาจจะไม่ช่วยบอกความหมายของบทความได้ หรือใช้แทนเนื้อหาของบทความได้ไม่ดีเพราะคำประเภทนี้จะปรากฏอยู่ในบทความเกือบทุกบทความอย่างเช่น คำภาษาอังกฤษ "is", "a", "this", "that" คำภาษาไทย "มี" "ใช้" "เป็น" ฯลฯ

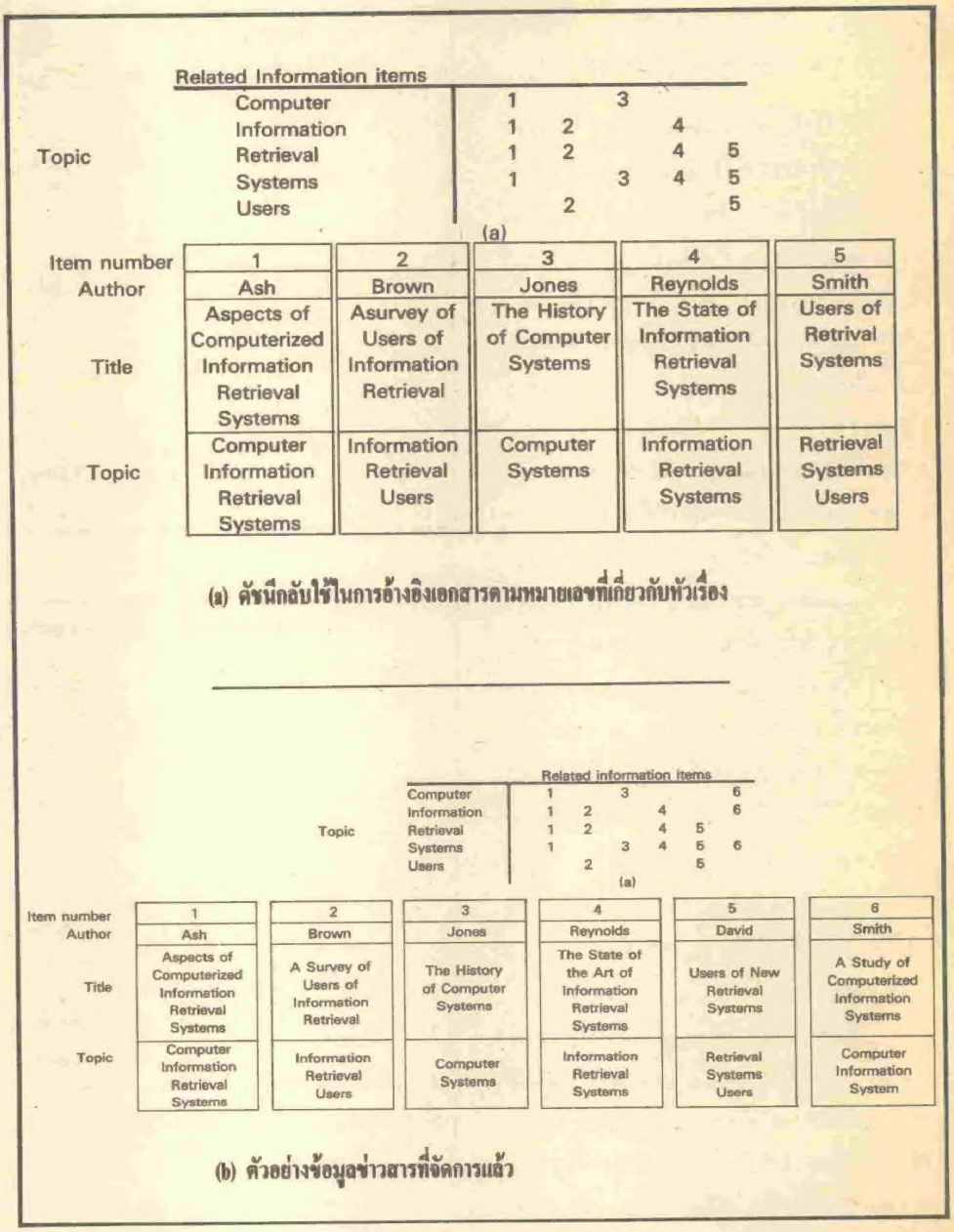
คำเหล่านี้จะถูกตัดออก การเลือกคำดัชนีนั้นมีหลักเกณฑ์คือควรจะเลือกคำที่มีความถี่ (จำนวนครั้งที่ปรากฏในบทความ) ปานกลางถึงความถี่ค่อนข้างต่ำ ซึ่งการเลือกคำดัชนีนี้จะมีผลต่อการค้นหาเอกสารด้วย เมื่อเลือกคำดัชนีที่เหมาะสมแล้วก็จะเป็นการจัดเก็บและสร้างโครงสร้าง

ในขั้นตอนการทำงานข้างต้นสำหรับเอกสารที่เป็นภาษาอังกฤษนั้นทำได้ไม่ยากเพราะคำภาษาอังกฤษในบทความได้ถูกแยกไว้อย่างเด่นชัด โดยมีช่องว่างคั่นระหว่างคำแต่ละคำ ส่วนภาษาไทยการจัดรูปแบบจะต้องคำนึงถึงการแยกคำออกจากประโยคด้วย ซึ่งต้องเพิ่มขั้นตอนในการทำงานขึ้นอีก ซึ่งเป็นที่น่ายินดีที่ในขณะนี้ได้มีผู้พัฒนาระบบแยกคำภาษาไทยแล้ว (ดูรามา Spell Check ไมโครคอมพิวเตอร์ ฉบับที่ 86)

## จัดโครงสร้าง

หลังจากที่จัดรูปแบบและทำดัชนีเรียบร้อยแล้ว เอกสารและคำดัชนี (Index Word) ก็จะถูกจัดเก็บให้เป็นโครงสร้างที่สามารถใช้ในการค้นหาได้ เทคนิคในการจัดเก็บและสร้างโครงสร้างของข้อมูลข่าวสารมีด้วยกันสองวิธีหลัก ๆ วิธีแรกคือ การใช้ Inverted files System หรือแฟ้มเชิงกลับ ซึ่งเป็นเทคนิคที่ใช้กันในระบบฐานข้อมูล การค้นหาเอกสารข้อมูลที่ต้องการนั้นก็จะยึดหลักการจับคู่เหมือน (Matching) ระหว่างคำที่เป็นหัวข้อเรื่องเกี่ยวกับเรื่องหรือบทความที่เราต้องการ (Query terms) กับคำที่ถูกกำหนดให้เป็นคำดัชนีของเอกสารข้อมูล (Document) นั้น ๆ การทำแมตชิ่งนี้ทำได้ โดยใช้หลักตรรกศาสตร์บูลลีเอน (Boolean logic คือ AND, (OR และ NOT) ดังรูปที่ 1 บริษัทที่มีผลิตภัณฑ์ Inverted file System คือ Fulcrum Technologies

การจัดเก็บโครงสร้างของข้อมูลข่าวสาร อีกแบบหนึ่งก็คือ Vector System การจัดแบบนี้คำดัชนีจะถูกรวมให้เป็นชุด Vector และเอกสารหนึ่ง ๆ จะถูกกำหนดให้เวกเตอร์ของคำดัชนีเป็นตัวแทน ซึ่งเวกเตอร์นี้จะมีน้ำหนัก (weight) แสดงค่าของคำดัชนีว่ามีความสัมพันธ์กับเอกสารนั้นมากน้อยเท่าไหร่ (น้ำหนักอาจมีค่าระหว่าง 0.0 ถึง 1.0 น้ำหนัก 0.0 หมายความว่า คำดัชนีนั้นไม่มีค่าความสัมพันธ์กับเอกสารเลย



Related Information items

| Topic | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Computer | 1 | | 3 | | | |
| Information | 1 | 2 | | 4 | | |
| Retrieval | 1 | 2 | | 4 | 5 | |
| Systems | 1 | | 3 | 4 | 5 | |
| Users | | 2 | | | 5 | |

(a)

| Item number / Author | 1 Ash | 2 Brown | 3 Jones | 4 Reynolds | 5 Smith |
|---|---|---|---|---|---|
| Title | Aspects of Computerized Information Retrieval Systems | Asurvey of Users of Information Retrieval | The History of Computer Systems | The State of Information Retrieval Systems | Users of Retrival Systems |
| Topic | Computer Information Retrieval Systems | Information Retrieval Users | Computer Systems | Information Retrieval Systems | Retrieval Systems Users |

(a) ดัชนีกลับใช้ในการอ้างอิงเอกสารตามหมายเลขที่เกี่ยวกับหัวเรื่อง

Related information items

| Topic | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Computer | 1 | | 3 | | | 6 |
| Information | 1 | 2 | | 4 | | 6 |
| Retrieval | 1 | 2 | | 4 | 5 | |
| Systems | 1 | | 3 | 4 | 5 | 6 |
| Users | | 2 | | | 5 | |

(a)

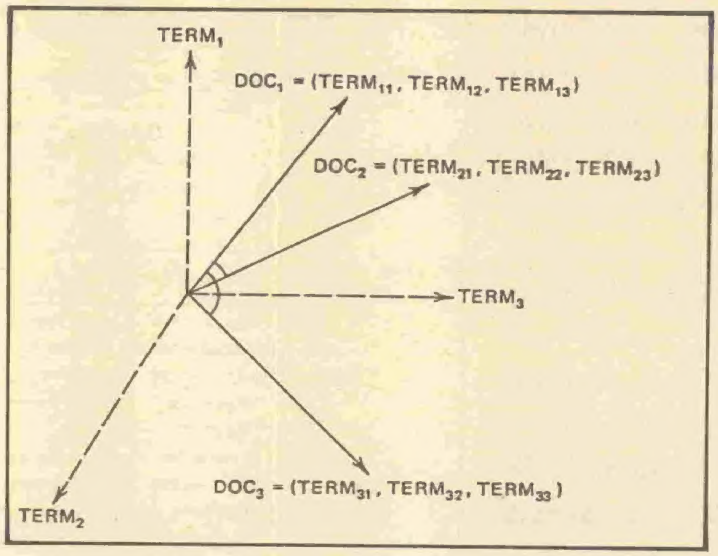
| Item number / Author | 1 Ash | 2 Brown | 3 Jones | 4 Reynolds | 5 David | 6 Smith |
|---|---|---|---|---|---|---|
| Title | Aspects of Computerized Information Retrieval Systems | A Survey of Users of Information Retrieval | The History of Computer Systems | The State of the Art of Information Retrieval Systems | Users of New Retrieval Systems | A Study of Computerized Information Systems |
| Topic | Computer Information Retrieval Systems | Information Retrieval Users | Computer Systems | Information Retrieval Systems | Retrieval Systems Users | Computer Information System |

(b) ตัวอย่างข้อมูลข่าวสารที่จัดการแล้ว

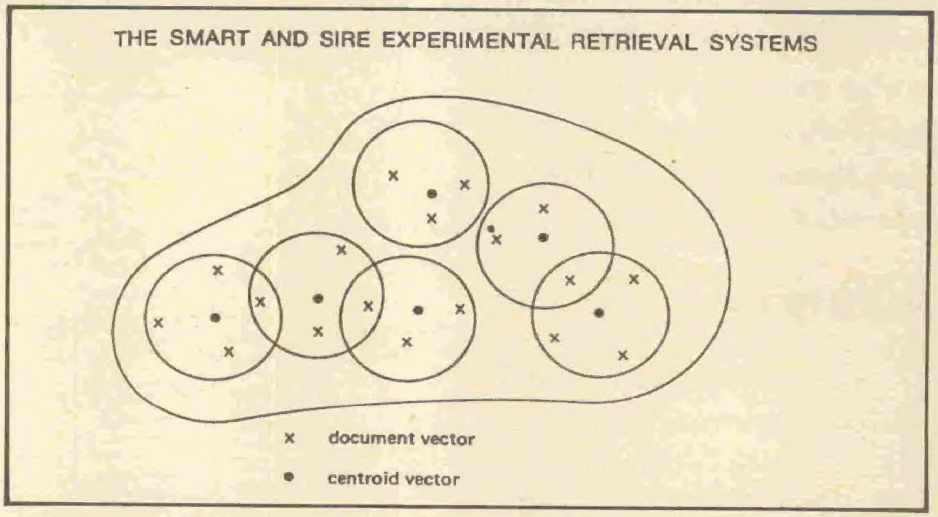
รูปที่ 1 การจัดโครงสร้าง Inverted file แบบง่าย ๆ

น้ำหนัก 1.0 หมายความว่า คำดัชนีนั้นมีความสัมพันธ์กับเอกสารมาก)

สำหรับการค้นหาเอกสารที่เราต้องการนั้นกระทำได้โดยการคำนวณหาความคล้ายคลึง (similarity) กันระหว่างเวกเตอร์ตัวแทนสิ่งที่ผู้ค้นหาต้องการ (query vector) กับเวกเตอร์ตัวแทนเอกสารแต่ละเอกสารดังรูปที่ 2 ด้วยการวัดความคล้ายคลึงนี้เองเราสามารถจัดแต่ละเอกสารให้อยู่รวมกันเป็นหมวดหมู่ (clusted document collection) เทคนิคการจัดแบบนี้ช่วยให้การค้นหาเอกสารง่ายขึ้นและเร็วขึ้น เพราะไม่ต้องคำนวณค่าความคล้ายคลึงของเวกเตอร์ค้นหา (query vector) กับเวกเตอร์แทนเอกสารทุก ๆ ชิ้นแต่จะคำนวณกับจุดศูนย์กลางของหมวดเอกสารนั้น ๆ (Centroid Vector) แทน ดังรูปที่ 3 ผู้ที่ผลิตระบบเวกเตอร์ก็คือ Thinking Machines ผู้ผลิต Connection Machine




รูปที่ 2
การจัดเก็บโครงสร้างข้อมูลข่าวสารแบบเวกเตอร์




รูปที่ 3 Clustered Document Collection

## ประเมินผล

ในการค้นหาเอกสารข้อมูลที่เราต้องการ (relevant document) ผู้ค้นมีความต้องการให้ระบบข้อมูลข่าวสารเรียกเอกสารที่มีเนื้อหาเกี่ยวข้องกับความต้องการของผู้ค้น (Recall) หรือความครบถ้วนให้ได้เป็นจำนวนมาก ๆ และต้องการให้ระบบข่าวสารเรียกเอกสารที่เนื้อหาใกล้เคียงกับความต้องการมาก ๆ (Precision) หรือความชัดเจน

ระบบข้อมูลข่าวสารนั้นถูกพัฒนาขึ้นเพื่อให้ทั้งความครบถ้วนและความชัดเจนมีค่ามากที่สุด แต่ในความเป็นจริงหากระบบเรียกเอกสารได้ครบถ้วน (ไม่ทิ้งเอกสารที่เกี่ยวข้องเลย) ความชัดเจนก็จะน้อยลง

## ข้อมูลข่าวสารในวันนี้

ในยุคที่ข่าวสารข้อมูลเพิ่มจำนวนขึ้นไม่เว้นแต่ละวันนี้ มีข่าวสารข้อมูลที่มีประโยชน์ไม่น้อยที่ไม่สามารถนำมาใช้ประโยชน์ได้ ทั้งนี้ก็เนื่องมาจากการค้นหาข้อมูลข่าวสารนั้นกระทำได้ยากและใช้เวลานาน หากมีการนำเอาระบบข้อมูลข่าวสารเข้ามาใช้ก็จะทำให้ค้นหาข้อมูลข่าวสารที่ต้องการได้ง่าย และเป็นประโยชน์ต่อการค้นคว้าหาความรู้ด้วย

ระบบข้อมูลข่าวสารนี้สามารถนำไปใช้ในการเก็บข้อมูลเอกสาร หนังสือในห้องสมุด เพื่อให้ผู้ใช้ห้องสมุดได้ค้นหาข้อมูล หนังสือที่ต้องการได้ง่ายและรวดเร็วขึ้น หรืออาจจะนำเข้าไปใช้ในการเก็บข้อมูลบทความ ข้อมูลของสิ่งพิมพ์ หนังสือพิมพ์ วารสารต่าง ๆ เพื่อง่ายต่อการค้นหาและสามารถทำเป็นการค้าในการขายข้อมูลข่าวสารแก่ผู้สนใจได้อีกด้วย เนื่องจากการเพิ่มจำนวนของข้อมูลข่าวสารนี้เอง ทั้งข้อมูลข่าวสารที่เราต้องการและไม่ต้องการก็เพิ่มขึ้นเป็นจำนวนมาก หากเราไม่มีวิธีการค้นหาข้อมูลที่เราต้องการอย่างชาญฉลาดและรวดเร็วเราจะไม่สามารถได้รับประโยชน์จากข้อมูลข่าวสารที่เพิ่มขึ้นได้ ถ้าเราต้องเสียเวลาอ่านเอกสารที่ไม่เกี่ยวข้องกับความต้องการ และไม่อาจค้นพบสิ่งที่เราต้องการได้ ระบบจัดเก็บและค้นหาข้อมูลข่าวสารจึงมีความสำคัญในการจัดการกับข้อมูลข่าวสารที่มีอยู่เป็นจำนวนมากนี้ เพื่อให้ได้รับประโยชน์จากข้อมูลข่าวสารที่มีอยู่ให้มากที่สุด

## อนาคตของระบบ

การเพิ่มจำนวนของข้อมูลข่าวสารอย่างไม่มีที่สิ้นสุดนี้ เราจะเห็นได้บางครั้งเสียเวลาไปมากกับการรีวิวเอกสารบทความบางฉบับที่ไม่ตรงกับความต้องการของเรา แม้ว่าเราจะขยันอ่านสักเพียงใดก็ตาม ก็ไม่สามารถติดตามอ่านข้อมูลข่าวสารได้ทั้งหมด การนำระบบจัดเก็บและค้นหาข้อมูลข่าวสารนี้เข้ามาช่วยทำงาน จะทำให้การทำงานของเราเป็นไปอย่างมีประสิทธิภาพและรวดเร็วยิ่งขึ้น เพื่อให้ทันต่อเหตุการณ์ในโลกแห่งการแข่งขัน ระบบข้อมูลข่าวสารในอนาคตจะเข้ามามีบทบาทเป็นผู้กลั่นกรองข้อมูลข่าวสารและเลือกหาข้อมูลที่เป็นประโยชน์ต่อผู้ใช้ หรือเรียกได้ว่าเป็นหูเป็นตาแทนเลยก็ได้ เราน่าจะจับตามองว่าเมื่อไหร่จะนำระบบข้อมูลข่าวสารที่เป็นภาษาไทยมาใช้ในบ้านเรากันสักที

หากเราต้องการเอกสารที่เกี่ยวกับ information Q = Information ระบบจะเลือกเอกสาร 1, 2 และ 4 หากเราต้องการเอกสารที่เกี่ยวกับ Computer และ Systems Q = Computer AND System ระบบเลือกเอกสาร 1 และ 3

พบแล้ว

สร้างภาพกราฟิก
และเอฟเฟ็กต์โดยเร็ว
การอัพเกรด CoreDRAW
จากเวอร์ชันก่อน ๆ

พลาดไม่ได้

หนังสือ เริ่มต้นใช้งาน CorelDRAW 3 เป็นหนังสือสองเล่มในเล่มเดียวกัน เป็นหนังสือแนะนำ CorelDRAW เวอร์ชันล่าสุด ประกอบด้วยตัวอย่างบทเรียนและหลักฐานอ้างอิงคำสั่งต่าง ๆ ที่ใช้ในโปรแกรมนี้อย่างครบถ้วนสมบูรณ์ ไม่ว่าท่านผู้อ่านจะยังใหม่ต่อโปรแกรม CorelDRAW หรือว่าต้องการจะอัพเกรด CorelDRAW จากเวอร์ชันก่อน ๆ ก็ตาม

หาซื้อได้ตามร้านหนังสือชั้นนำทั่วไป
หรือที่ ซีเอ็ดบุ๊คเซ็นเตอร์ทุกสาขา
และสั่งซื้อได้ที่ซีเอ็ด สำนักงานใหญ่
โทร 248-6280

มีจำหน่ายแล้ว